Machine Learning with Python

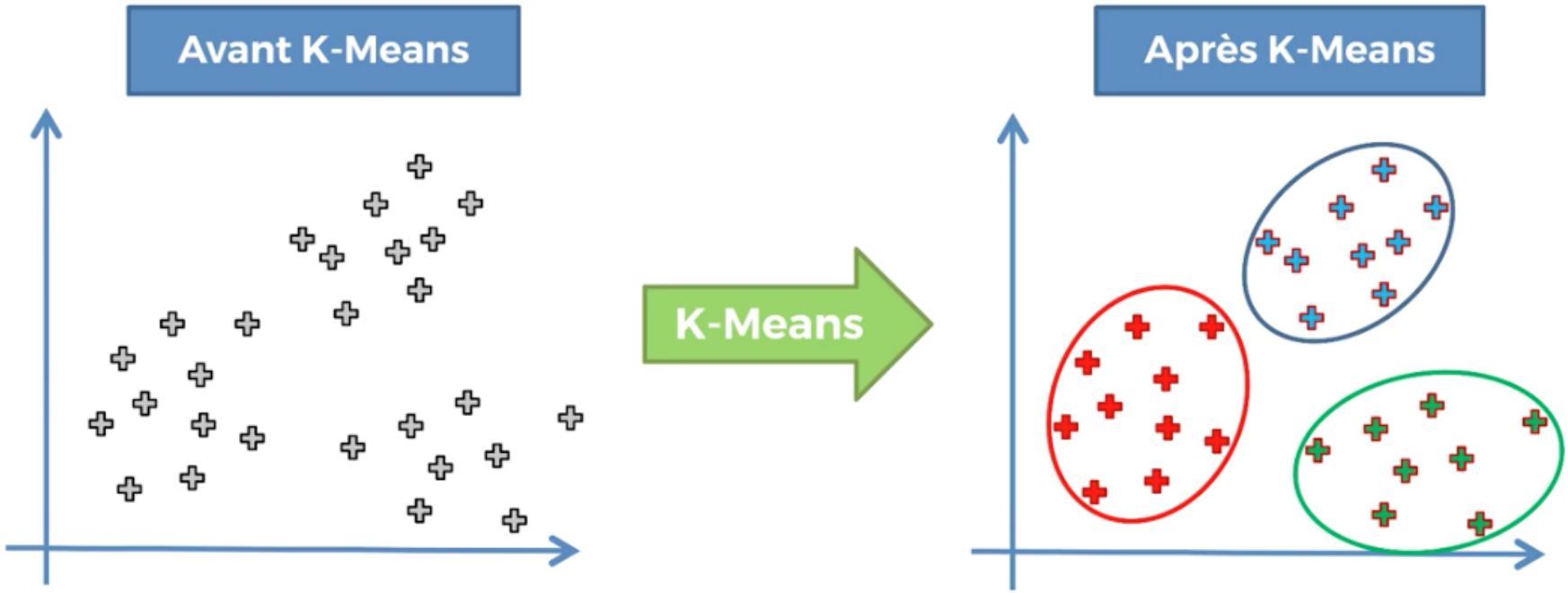
K-means clustering

Bassem Ben Hamed

Juillet 2018

K-Means Intuition: Comprendre K-Means

Que fait K-Means ?



L'algorithme K-Means

STEP 1: Choisir le nombre K de clusters

STEP 2: Sélectionner au hasard K points, les centroids

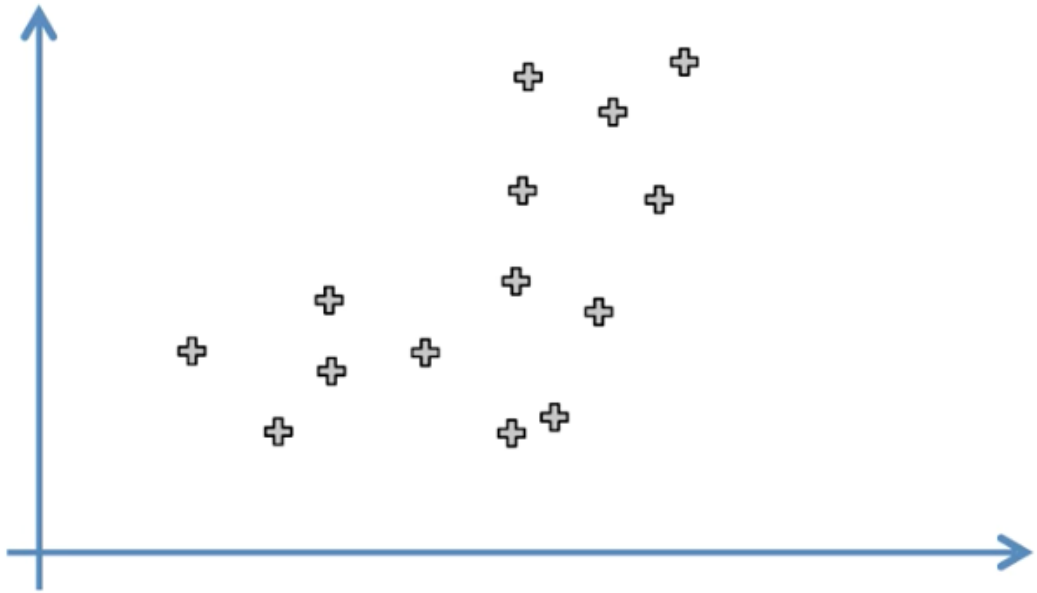
STEP 3: Assigner chaque point au centroid le plus proche ➡ Cela forme K clusters

STEP 4: Calculer et placer le nouveau centroid de chaque cluster

STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon: **Votre modèle est prêt**

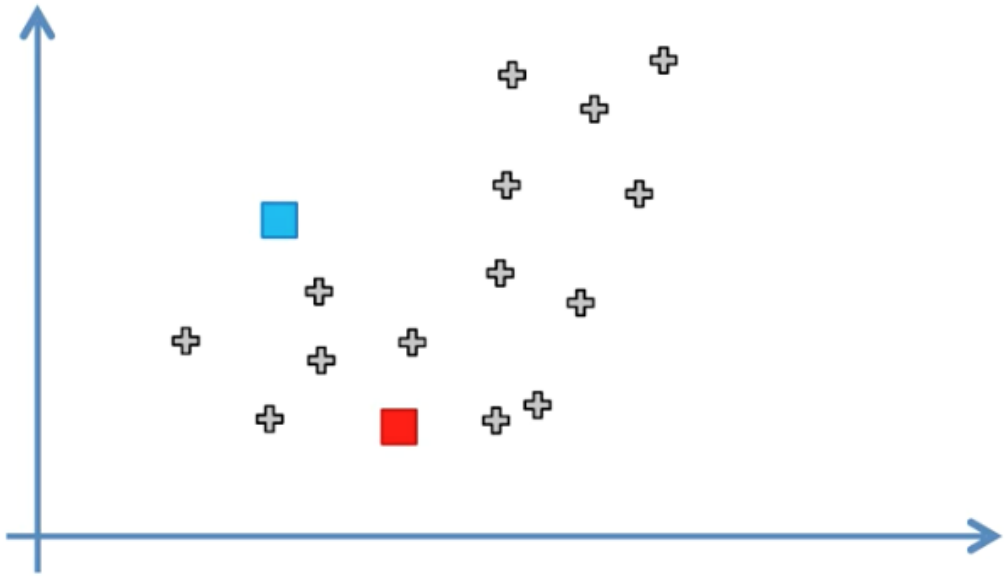
L'algorithme K-Means

STEP 1: Choisir le nombre K de clusters: K = 2



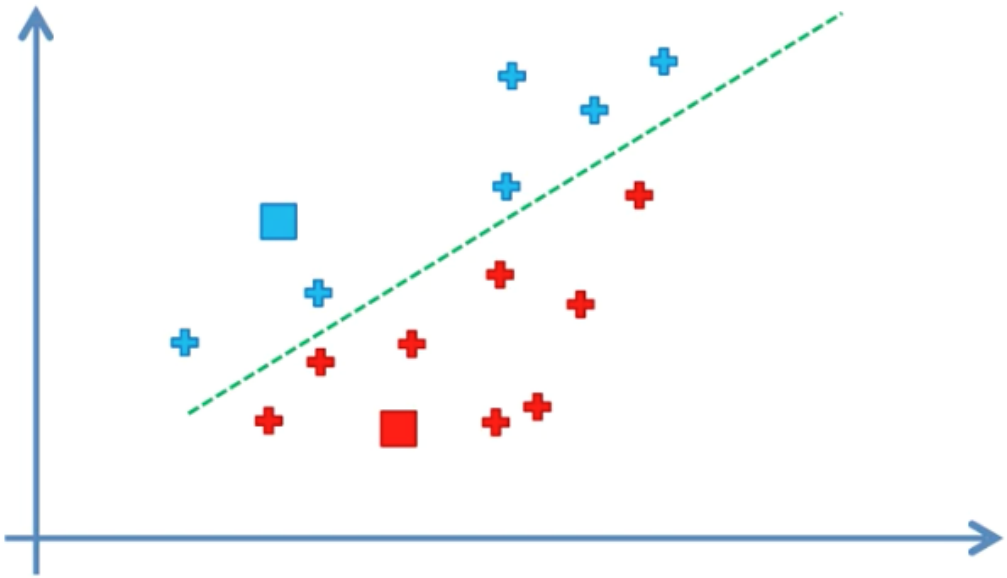
L'algorithme K-Means

STEP 2: Sélectionner au hasard K points, les centroids (pas nécessairement du dataset)



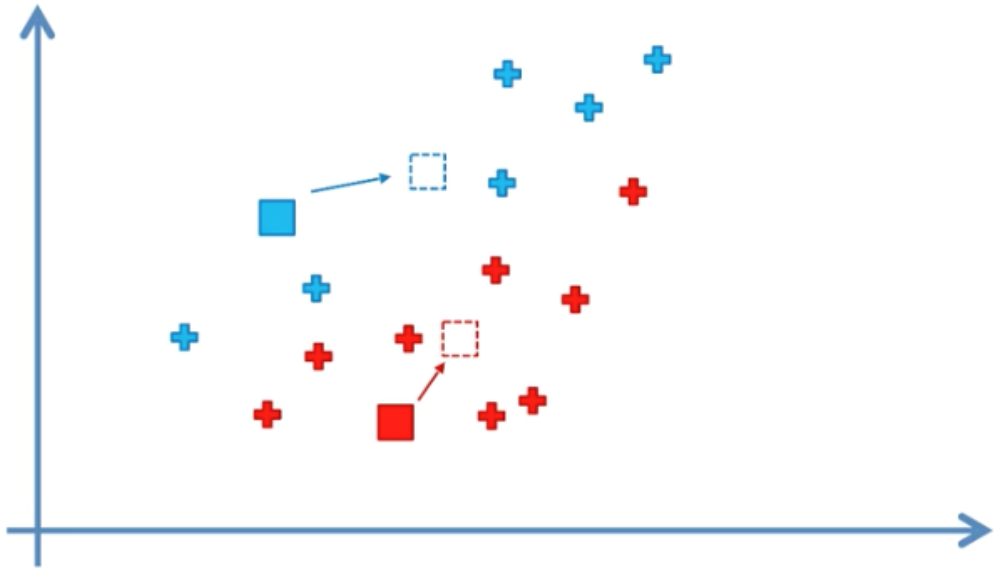
L'algorithme K-Means

STEP 3: Assigner chaque point au centroid le plus proche ➡ Cela forme K clusters



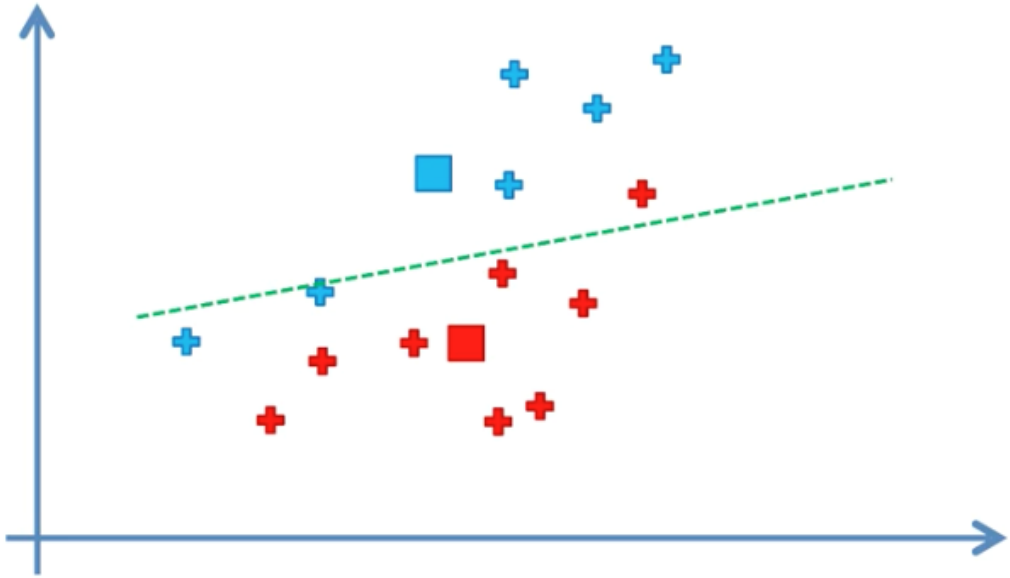
L'algorithme K-Means

STEP 4: Calculer et placer le nouveau centroid de chaque cluster



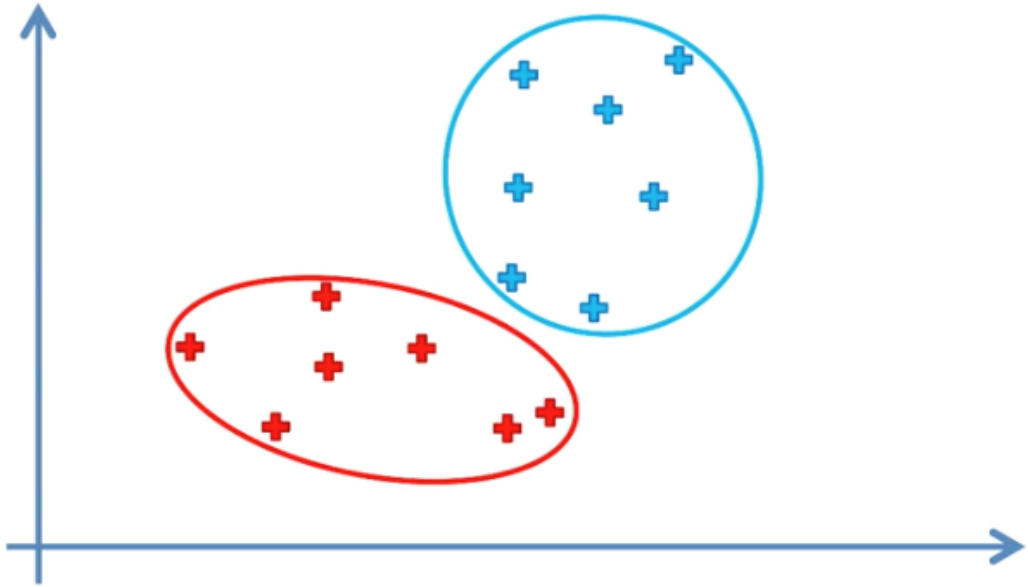
L'algorithme K-Means

STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



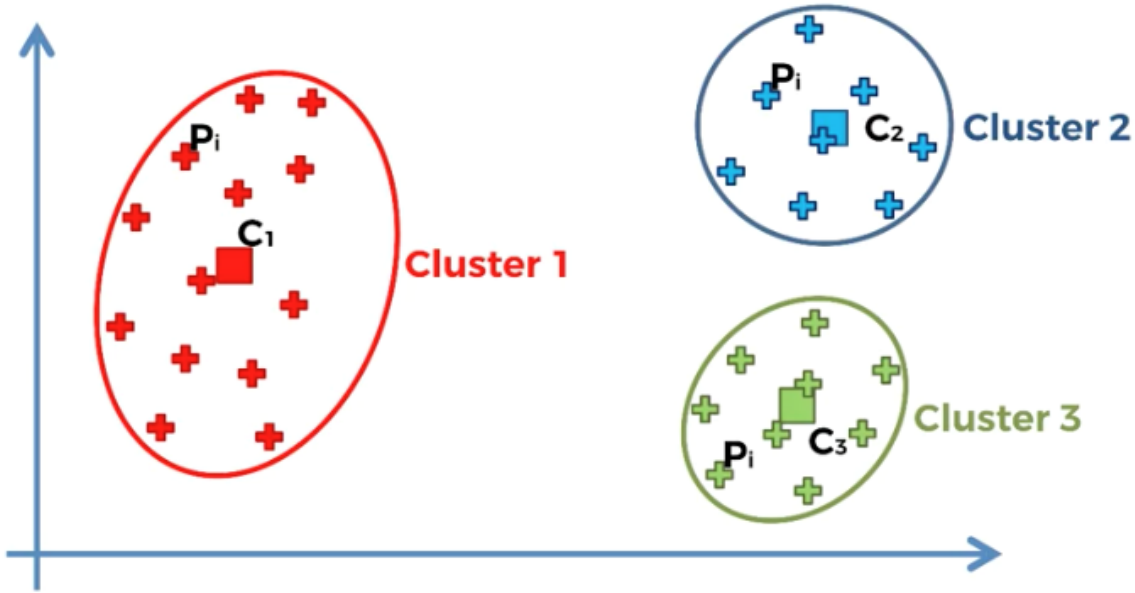
L'algorithme K-Means

FIN: Votre modèle est prêt



K-Means Intuition:
Choisir le bon nombre de clusters

Choisir le bon nombre de clusters

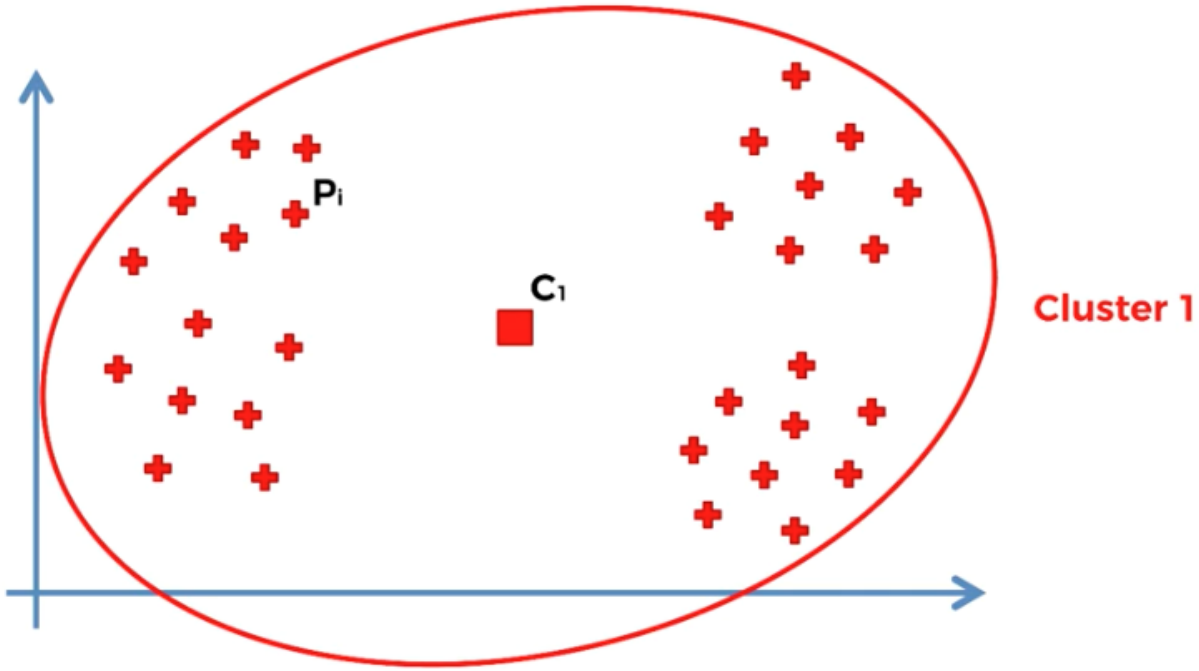


$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Choisir le bon nombre de clusters

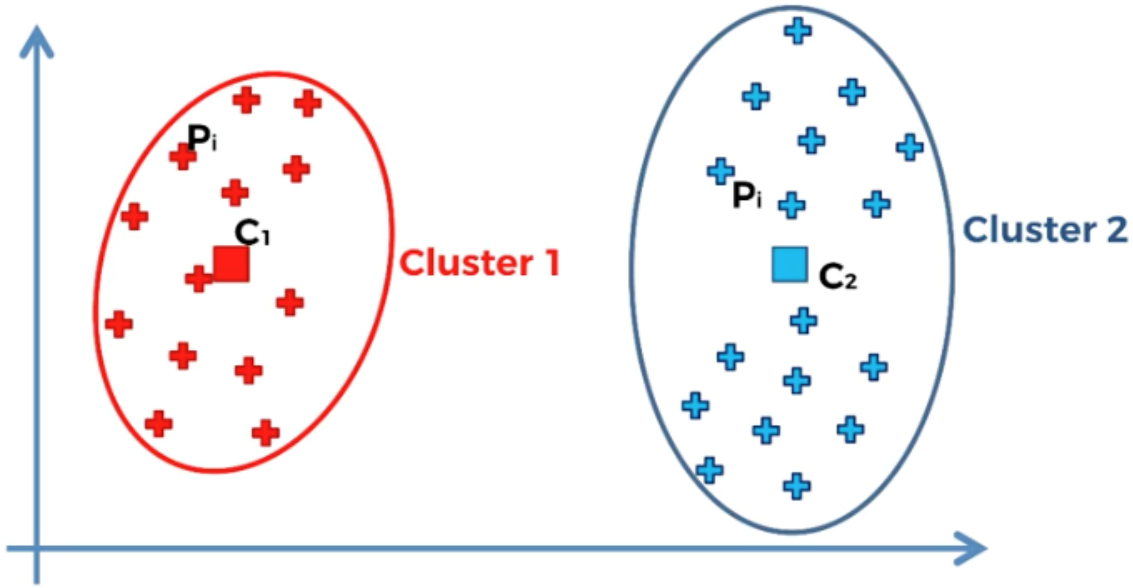
Rembobinons...

Choisir le bon nombre de clusters



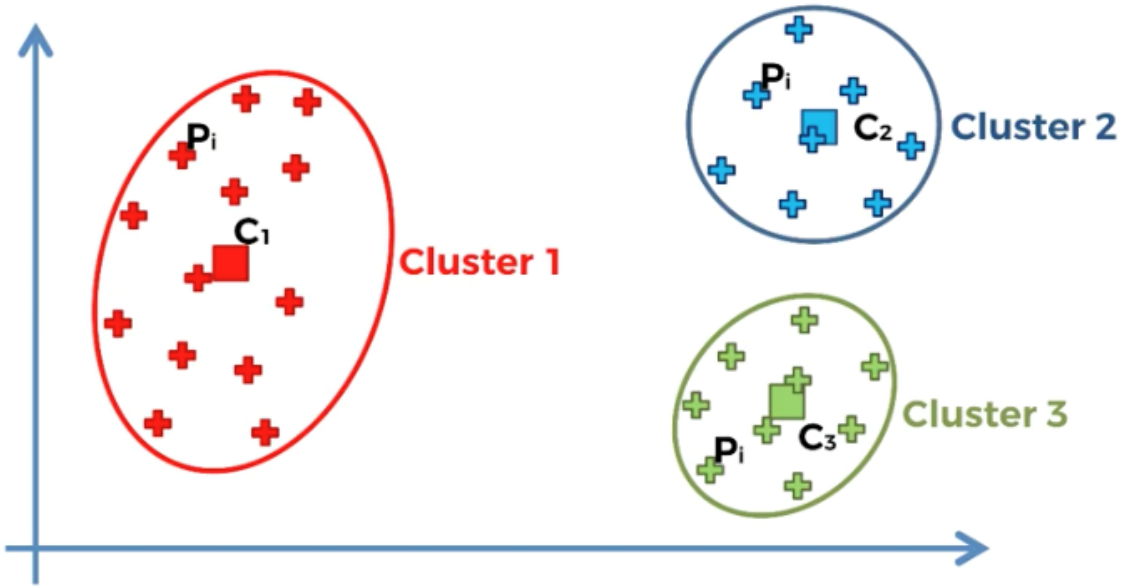
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2$$

Choisir le bon nombre de clusters



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2$$

Choisir le bon nombre de clusters



$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Choisir le bon nombre de clusters